

# Weekly Report

December 10, 2017

## 1 Work

### 1.1 降维

本周都在进行降维的工作，主要完成了以下工作：

- 开始使用kmeans的评价函数：当降维点的移动开始稳定的时候开始使用Kmeans。
- 停止基于团移动的评价函数：当团的位移开始稳定的时候，转为基于点的优化。
- 性能评价指标，1NN classifier。
- 目前已经将结果运用于数据集：MNIST(70,00\*784,10类)，fashion MNIST(70,000\*784,10类)，Twitter的word2vec结果(600,000\*784,10类)。

Table 1: 实验结果

100NN Graph	KNN Construction	Embedding	Kmeans	Total	1NN Classifier Accuracy
MNIST	40s(99.95%准确率)	37s	2s	77s	93.9457%
Fashion MNIST	37s(98.71%准确率)	38s	2s	75s	72.9029%
Twitter	175s(76.80%准确率)	282s	108s	457s	56.1362%

使用kmeans的方法可以加速相似团的合并，否则在后期属于同一个类的两个团难以合并。图1(a)是使用kmeans得到的结果，图1(b)是不使用kmeans得到的结果，他们的总迭代次数一致。

同时我们发现（bhtnse的作者也提到了这个小trick），一开始增加邻近点的相似性，能够提前让相似的团聚在一起，从而避免在最后相似团不能靠拢的问题。在实践中，我们在迭代的前10%，降低不相似点的排斥性（也就是增加邻近

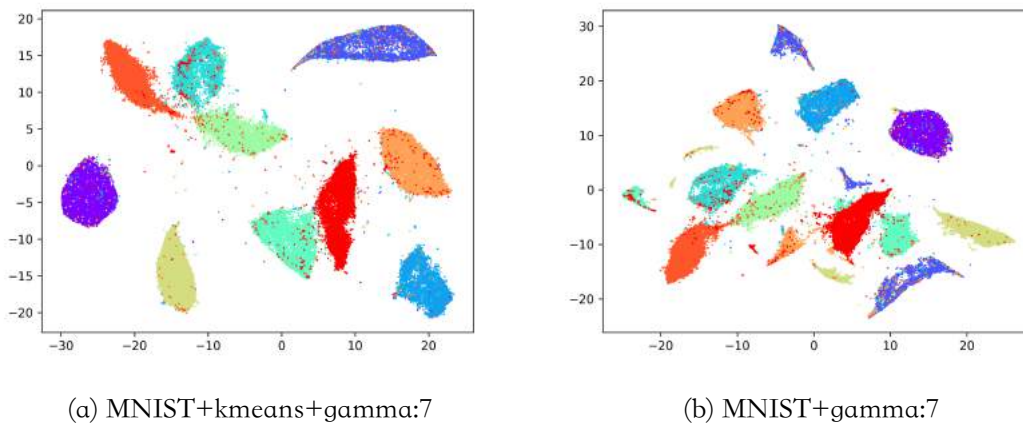


Figure 1

点的相似性），这样点之间的吸引力就占据主要成分，我们可以获得更好的投影。图2显示的是在迭代过程10%的时候，(a)是增加邻近点相似性的结果，相似的点提前聚合在一起。

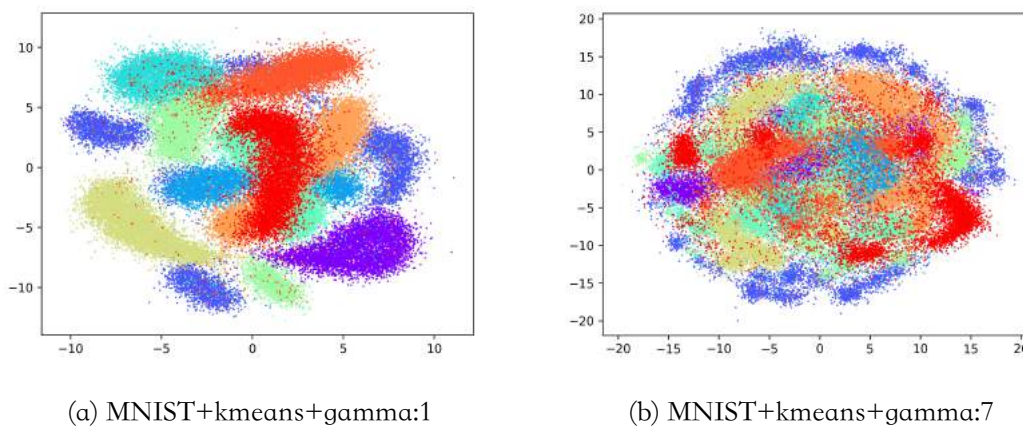


Figure 2

在使用Twitter数据的过程中，由于数据是对Twitter的word2vec的学习结果，没有label标签，我们在原始空间中使用Kmeans聚10类作为标签。也由于这类数据本身不带有聚类的内部结构，所以投影过程中大部分点仍然聚集在一起，所以Kmeans的作用不是很明显（图3）。

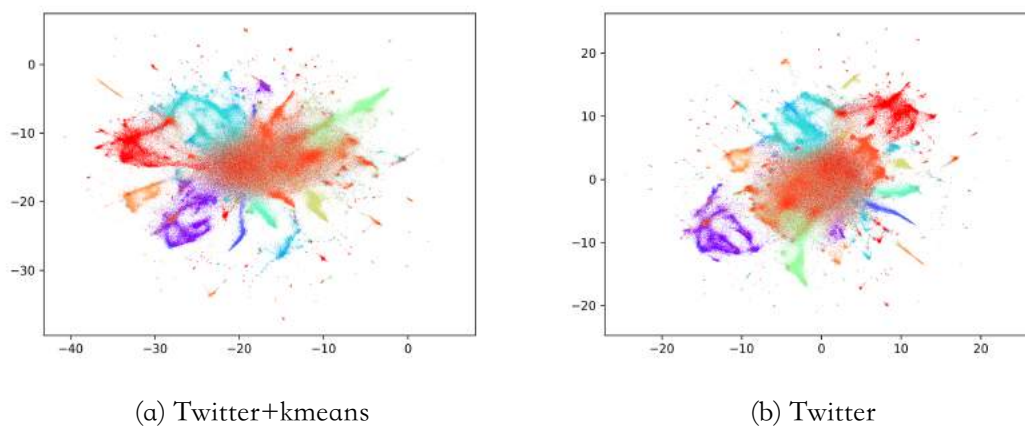


Figure 3

Table 2: 工作进度

TASK	PROGRESS	DATE
dimension reduction	已经基本完成整体框架，后期需要1) 代码重构一下，节省内存2) 调整参数3) 运行更大的数据。	12.30
location2vec专利	约好下周一和律师联系	
*2Vec survey		1.30

## 1.2 工作进度

## 2 Paper Reading

### 2.1 SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives

在优化过程中，有许多更新迭代学习参数的方法，SAGA能够保持优化方面无偏的情况下，在理论上获得更快的迭代速度。同时，SAGA对目标函数的要求降低为非强凸，可以对更多目标函数做处理。